Introduction of Markov Decision Process

Prof. John C.S. Lui
Department of Computer Science & Engineering
The Chinese University of Hong Kong

Motivation

Why Markov Decision Process?

- To decide on a proper (or optimal) policy.
- To maximize performance measures.
- To obtain transient measures.
- To obtain long-term measures (fixed or discounted).
- To decide on the *optimal* policy via an efficient method (using dynamic programming).

Review of DTMC

Toymaker

- A toymaker is involved in a toy business.
- Two states: state 1 is toy is favorable by public, state 2 otherwise.
- State transition (per week) is:

$$\boldsymbol{P} = [p_{ij}] = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{2}{5} & \frac{3}{5} \end{bmatrix}$$

- What is the *transient* measure, say state probability?

Transient State Probability Vector

Transient calculation

Assume the MC has N states.
Let $\pi_i(n)$ be the probability of system at state i *after* n transitions if its state at $n = 0$ is known.
We have:

$$\begin{aligned} \sum_{i=1}^{N} \pi_i(n) &= 1 && (1) \\ \pi_j(n+1) &= \sum_{i=1}^{N} \pi_i(n) p_{ij} \ \text{ for } n = 0, 1, 2, .. && (2) \end{aligned}$$

Transient State Probability Vector

Iterative method

In vector form, we have:

$$\pi(n+1) = \pi(n)\boldsymbol{P} \;\; \text{for } n = 0, 1, 2, ...$$

or

$$\begin{aligned}
\pi(1) &= \pi(0)\boldsymbol{P} \\
\pi(2) &= \pi(1)\boldsymbol{P} = \pi(0)\boldsymbol{P}^2 \\
\pi(3) &= \pi(2)\boldsymbol{P} = \pi(0)\boldsymbol{P}^3 \\
\ldots & \quad \ldots \\
\pi(n) &= \pi(0)\boldsymbol{P}^n \;\; \text{for } n = 0, 1, 2, ...
\end{aligned}$$

Illustration of toymaker

Assume $\pi(0) = [1, 0]$

$n =$	0	1	2	3	4	5	...
$\pi_1(n)$	1	0.5	0.45	0.445	0.4445	0.44445	....
$\pi_2(n)$	0	0.5	0.55	0.555	0.5555	0.55555	....

Assume $\pi(0) = [0, 1]$

$n =$	0	1	2	3	4	5	...
$\pi_1(n)$	0	0.4	0.44	0.444	0.4444	0.44444	....
$\pi_2(n)$	1	0.6	0.56	0.556	0.5556	0.55556	....

Note π at steady state is *independent* of the initial state vector.

Review of z-transform

Examples:

Time Sequence $f(n)$	**z-transform** $F(z)$
$f(n) = 1$ if $n \geq 0$, 0 otherwise	$\frac{1}{1-z}$
$kf(n)$	$kF(z)$
$\alpha^n f(n)$	$F(\alpha z)$
$f(n) = \alpha^n$, for $n \geq 0$	$\frac{1}{1-\alpha z}$
$f(n) = n\alpha^n$, for $n \geq 0$	$\frac{\alpha z}{(1-\alpha z)^2}$
$f(n) = n$, for $n \geq 0$	$\frac{z}{(1-z)^2}$
$f(n-1)$, or shift left by one	$zF(z)$
$f(n+1)$, or shift right by one	$z^{-1}\left[F(z) - f(0)\right]$

z-transform of iterative equation

$$\pi(n+1) = \pi(n)\boldsymbol{P} \quad \text{for } n = 0, 1, 2, ...$$

Taking the z-transform:

$$\begin{aligned} z^{-1}\left[\boldsymbol{\Pi}(z) - \pi(0)\right] &= \boldsymbol{\Pi}(z)\boldsymbol{P} \\ \boldsymbol{\Pi}(z) - z\boldsymbol{\Pi}(z)\boldsymbol{P} &= \pi(0) \\ \boldsymbol{\Pi}(z)\left(\boldsymbol{I} - z\boldsymbol{P}\right) &= \pi(0) \\ \boldsymbol{\Pi}(z) &= \pi(0)\left(\boldsymbol{I} - z\boldsymbol{P}\right)^{-1} \end{aligned}$$

We have $\boldsymbol{\Pi}(z) \Leftrightarrow \pi(n)$ and $(\boldsymbol{I} - z\boldsymbol{P})^{-1} \Leftrightarrow \boldsymbol{P}^n$. In other words, from $\boldsymbol{\Pi}(z)$, we can perform transform inversion to obtain $\pi(n)$, for $n \geq 0$, which gives us the transient probability vector.

Example: Toymaker

Given:

$$\boldsymbol{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \\ \frac{2}{5} & \frac{3}{5} \end{bmatrix}$$

We have:

$$(\boldsymbol{I} - z\boldsymbol{P}) = \begin{bmatrix} 1 - \frac{1}{2}z & -\frac{1}{2}z \\ -\frac{2}{5}z & 1 - \frac{3}{5}z \end{bmatrix}$$

$$(\boldsymbol{I} - z\boldsymbol{P})^{-1} = \begin{bmatrix} \frac{1-\frac{3}{5}z}{(1-z)(1-\frac{1}{10}z)} & \frac{\frac{1}{2}z}{(1-z)(1-\frac{1}{10}z)} \\ \frac{\frac{2}{5}z}{(1-z)(1-\frac{1}{10}z)} & \frac{1-\frac{1}{2}z}{(1-z)(1-\frac{1}{10}z)} \end{bmatrix}$$

$$
(\boldsymbol{I} - z\boldsymbol{P})^{-1} = \left[\begin{array}{cc} \frac{4/9}{1-z} + \frac{5/9}{1-\frac{z}{10}} & \frac{5/9}{1-z} + \frac{-5/9}{1-\frac{z}{10}} \\ \\ \frac{4/9}{1-z} + \frac{-4/9}{1-\frac{z}{10}} & \frac{5/9}{1-\frac{z}{10}} + \frac{4/9}{1-\frac{z}{10}} \end{array}\right]
$$

$$
= \frac{1}{1-z}\left[\begin{array}{cc} 4/9 & 5/9 \\ 4/9 & 5/9 \end{array}\right] + \frac{1}{1-\frac{1}{10}z}\left[\begin{array}{cc} 5/9 & -5/9 \\ -4/9 & 4/9 \end{array}\right]
$$

Let $\boldsymbol{H}(n)$ be the inverse of $(\boldsymbol{I} - z\boldsymbol{P})^{-1}$ (or $\boldsymbol{P}^n$):

$$
\boldsymbol{H}(n) = \left[\begin{array}{cc} 4/9 & 5/9 \\ 4/9 & 5/9 \end{array}\right] + \left(\frac{1}{10}\right)^n \left[\begin{array}{cc} 5/9 & -5/9 \\ -4/9 & 4/9 \end{array}\right] = \boldsymbol{S} + \boldsymbol{T}(n)
$$

Therefore:

$$
\pi(n) = \pi(0)\boldsymbol{H}(n) \quad \text{for } n = 0, 1, 2...
$$

A closer look into $\boldsymbol{P}^n$

What is the convergence rate of a particular MC? Consider:

$$\boldsymbol{P} = \begin{bmatrix} 0 & 3/4 & 1/4 \\ 1/4 & 0 & 3/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix},$$

$$(\boldsymbol{I} - z\boldsymbol{P}) = \begin{bmatrix} 1 & -\frac{3}{4}z & -\frac{1}{4}z \\ -\frac{1}{4}z & 1 & -\frac{3}{4}z \\ -\frac{1}{4}z & -\frac{1}{4}z & 1-\frac{1}{2}z \end{bmatrix}.$$

A closer look into $\boldsymbol{P}^n$: continue

We have

$$
\begin{aligned}
\det(\boldsymbol{I} - z\boldsymbol{P}) &= 1 - \frac{1}{2}z - \frac{7}{16}z^2 - \frac{1}{16}z^2 \\
&= (1 - z)\left(1 + \frac{1}{4}z\right)^2
\end{aligned}
$$

It is easy to see that $z = 1$ is always a root of the determinant for an irreducible Markov chain (which corresponds to the equilibrium solution).

A closer look into $\boldsymbol{P}^n$: continue

$$
\begin{aligned}
[\boldsymbol{I} - z\boldsymbol{P}]^{-1} &= \frac{1}{(1-z)[1+(1/4)z]^2} \\
&\times \begin{bmatrix}
1 - \frac{1}{2}z - \frac{3}{16}z^2 & \frac{3}{4}z - \frac{5}{16}z^2 & \frac{1}{4}z + \frac{9}{16}z^2 \\
\frac{1}{4}z - \frac{1}{16}z^2 & 1 - \frac{1}{2}z - \frac{1}{16}z^2 & \frac{3}{4}z + \frac{1}{16}z^2 \\
\frac{1}{4}z - \frac{1}{16}z^2 & 1 - \frac{1}{4}z - \frac{3}{16}z^2 & 1 - \frac{3}{16}z^2
\end{bmatrix}
\end{aligned}
$$

Now the only issue is to find the *inverse* via partial fraction expansion.

A closer look into $\boldsymbol{P}^n$: continue

$$
\begin{aligned}
[\boldsymbol{I} - z\boldsymbol{P}]^{-1} &= \frac{1/25}{1-z}\begin{bmatrix} 5 & 7 & 13 \\ 5 & 7 & 13 \\ 5 & 7 & 13 \end{bmatrix} + \frac{1/5}{(1+z/4)}\begin{bmatrix} 0 & -8 & 8 \\ 0 & 2 & -2 \\ 0 & 2 & -2 \end{bmatrix} \\
&+ \frac{1/25}{(1+z/4)^2}\begin{bmatrix} 20 & 33 & -53 \\ -5 & 8 & -3 \\ -5 & -17 & 22 \end{bmatrix}
\end{aligned}
$$

A closer look into P^n: continue

$$
\begin{aligned}
H(n) &= \frac{1}{25}\begin{bmatrix} 5 & 7 & 13 \\ 5 & 7 & 13 \\ 5 & 7 & 13 \end{bmatrix} + \frac{1}{5}(n+1)\left(-\frac{1}{4}\right)^n \begin{bmatrix} 0 & -8 & 8 \\ 0 & 2 & -2 \\ 0 & 2 & -2 \end{bmatrix} \\
&+\frac{1}{5}\left(-\frac{1}{4}\right)^n \begin{bmatrix} 20 & 33 & -53 \\ -5 & 8 & -3 \\ -5 & -17 & 22 \end{bmatrix} \quad n = 0, 1, \ldots
\end{aligned}
$$

A closer look into $\boldsymbol{P}^n$: continue

Important Points

- Equilibrium solution is *independent* of the initial state.
- Two *transient matrices*, which decay in the limit.
- The rate of decay is related to the *characteristic values*, which is one over the zeros of the determinant.
- The characteristic values are 1, $1/4$, and $1/4$.
- The decay rate at each step is $1/16$.

Motivation

- An $N-$state MC earns r_{ij} dollars when it makes a *transition* from state i to j.
- We can have a reward matrix $\boldsymbol{R} = [r_{ij}]$.
- The Markov process accumulates a sequence of rewards.
- What we want to find is the transient cumulative rewards, or even long-term cumulative rewards.
- For example, what is the *expected earning* of the toymaker in n weeks if he (she) is now in state i?

Let $v_i(n)$ be the expected total rewards in the next n transitions:

$$\begin{aligned} v_i(n) &= \sum_{j=1}^{N} p_{ij}[r_{ij} + v_j(n-1)] \quad i = 1, \ldots, N, \; n = 1, 2, ... && (3) \\ &= \sum_{j=1}^{N} p_{ij} r_{ij} + \sum_{j=1}^{N} p_{ij} v_j(n-1) \quad i = 1, \ldots, N, \; n = 1, 2, ... && (4) \end{aligned}$$

Let $q_i = \sum_{j=1}^{N} p_{ij} r_{ij}$, for $i = 1, \ldots, N$ and q_i is the expected reward for the next transition if the current state is i, and

$$v_i(n) = q_i + \sum_{j=1}^{N} p_{ij} v_j(n-1) \quad i = 1, \ldots, N, \; n = 1, 2, ... \qquad (5)$$

In vector form, we have:

$$\boldsymbol{v}(n) = \boldsymbol{q} + \boldsymbol{P}\boldsymbol{v}(n-1) \quad n = 1, 2, .. \qquad (6)$$

Example

Parameters

- Successful business and again a successful business in the following week, earns \$9.
- Unsuccessful business and again an unsuccessful business in the following week, loses \$7.
- Successful (or unsuccessful) business and an unsuccessful (successful) business in the following week, earns \$3.

Example

Parameters

- Reward matrix $\boldsymbol{R} = \begin{bmatrix} 9 & 3 \\ 3 & -7 \end{bmatrix}$, and $\boldsymbol{P} = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix}$.
- We have $\boldsymbol{q} = \begin{bmatrix} 0.5(9) + 0.5(3) \\ 0.4(3) + 0.6(-7) \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}$. Use:

$$v_i(n) = q_i + \sum_{j=1}^{N} p_{ij} v_j(n-1), \qquad \text{for } i = 1, 2, \; n = 1, 2, \ldots \tag{7}$$

- Assume $v_1(0) = v_2(0) = 0$, we have:

$n =$	0	1	2	3	4	5	...
$v_1(n)$	0	6	7.5	8.55	9.555	10.5555	....
$v_2(n)$	0	-3	-2.4	-1.44	-0.444	0.5556	....

Example

Observations

- If one day to go and if I am successful (unsuccessful), I should continue (stop) my business.
- If I am losing and I still have four or less days to go, I should stop.
- For large n, the long term average gain, $v_1(n) - v_2(n)$, has a difference of \$10 if I start from state 1 instead of state 2. In other words, starting from a successful business will have \$10 gain, as compare with an unsuccessful business.
- For large n, $v_1(n) - v_1(n-1) = 1$ and $v_2(n) - v_2(n-1) = 1$. In other words, each day brings a \$1 of profit.

$z-$transform reward analysis for toymaker

Equation (7) can be written:

$$v_i(n+1) = q_i + \sum_{j=1}^{N} p_{ij} v_j(n), \qquad \text{for } i = 1, 2,\ n = 0, 1, 2, \ldots$$

Apply $z-$transform, we have:

$$\begin{aligned}
z^{-1}\left[\boldsymbol{v}(z) - \boldsymbol{v}(0)\right] &= \frac{1}{1-z}\boldsymbol{q} + \boldsymbol{P}\boldsymbol{v}(z) \\
\boldsymbol{v}(z) - \boldsymbol{v}(0) &= \frac{z}{1-z}\boldsymbol{q} + z\boldsymbol{P}\boldsymbol{v}(z) \\
\left(\boldsymbol{I} - z\boldsymbol{P}\right)\boldsymbol{v}(z) &= \frac{z}{1-z}\boldsymbol{q} + \boldsymbol{v}(0) \\
\boldsymbol{v}(z) &= \frac{z}{1-z}\left(\boldsymbol{I} - z\boldsymbol{P}\right)^{-1}\boldsymbol{q} + \left(\boldsymbol{I} - z\boldsymbol{P}\right)^{-1}\boldsymbol{v}(0)
\end{aligned}$$

$z-$transform reward analysis for toymaker

Assume $\boldsymbol{v}(0) = \mathbf{0}$ (i.e., terminating cost is zero), we have:

$$\boldsymbol{v}(z) = \frac{z}{1-z}\left(\boldsymbol{I} - z\boldsymbol{P}\right)^{-1}\boldsymbol{q}. \tag{8}$$

Based on previous derivation:

$$(\boldsymbol{I} - z\boldsymbol{P})^{-1} = \frac{1}{1-z}\begin{bmatrix} 4/9 & 5/9 \\ 4/9 & 5/9 \end{bmatrix} + \frac{1}{1-\frac{1}{10}z}\begin{bmatrix} 5/9 & -5/9 \\ -4/9 & 4/9 \end{bmatrix}$$

$z-$transform reward analysis for toymaker

$$
\begin{aligned}
\frac{z}{1-z}(\boldsymbol{I}-z\boldsymbol{P})^{-1} &= \frac{z}{(1-z)^2}\begin{bmatrix} 4/9 & 5/9 \\ 4/9 & 5/9 \end{bmatrix} + \frac{z}{(1-z)(1-\frac{1}{10}z)}\begin{bmatrix} 5/9 & -5/9 \\ -4/9 & 4/9 \end{bmatrix} \\
&= \frac{z}{(1-z)^2}\begin{bmatrix} 4/9 & 5/9 \\ 4/9 & 5/9 \end{bmatrix} + \left(\frac{10/9}{1-z} + \frac{-10/9}{1-\frac{1}{10}z}\right)\begin{bmatrix} 5/9 & -5/9 \\ -4/9 & 4/9 \end{bmatrix}
\end{aligned}
$$

Let $\boldsymbol{F}(n) = [z/(1-z)]\,(\boldsymbol{I}-z\boldsymbol{P})^{-1}$, then

$$
\boldsymbol{F}(n) = n\begin{bmatrix} 4/9 & 5/9 \\ 4/9 & 5/9 \end{bmatrix} + \frac{10}{9}\left[1-\left(\frac{1}{10}\right)^n\right]\begin{bmatrix} 5/9 & -5/9 \\ -4/9 & 4/9 \end{bmatrix}
$$

Given that $\boldsymbol{q} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}$, we can obtain $\boldsymbol{v}(n)$ in closed form.

$z-$transform reward analysis for toymaker

$$\boldsymbol{v}(n) = n\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{10}{9}\left[1 - \left(\frac{1}{10}\right)^n\right]\begin{bmatrix} 5 \\ -4 \end{bmatrix} \qquad n = 0, 1, 2, 3...$$

When $n \rightarrow \infty$, we have:

$$v_1(n) = n + \frac{50}{9} \quad ; v_2(n) = n - \frac{40}{9}.$$

- For large n, $v_1(n) - v_2(n) = 10$.
- For large n, the slope of $v_1(n)$ or $v_2(n)$, the average reward per transition, is 1, or one unit of return per week. We can the average reward per transition the **gain**.

Asymptotic Behavior: for long duration process

- We derived this previously:

$$\boldsymbol{v}(z) = \frac{z}{1-z}\left(\boldsymbol{I} - z\boldsymbol{P}\right)^{-1}\boldsymbol{q} + \left(\boldsymbol{I} - z\boldsymbol{P}\right)^{-1}\boldsymbol{v}(0).$$

- The inverse transform of $(\boldsymbol{I} - z\boldsymbol{P})^{-1}$ has the form of $\boldsymbol{S} + \boldsymbol{T}(n)$.
- $\boldsymbol{S}$ is a stochastic matrix whose ith row is the limiting state probabilities if the system started in the ith state,
- $\boldsymbol{T}(n)$ is a set of differential matrices with geometrically decreasing coefficients.

Asymptotic Behavior: for long duration process

- We can write $(\boldsymbol{I} - z\boldsymbol{P})^{-1} = \frac{1}{1-z}\boldsymbol{S} + \mathcal{T}(z)$ where $\mathcal{T}(z)$ is the z-transform of $\boldsymbol{T}(n)$. Now we have

$$\boldsymbol{v}(z) = \frac{z}{(1-z)^2}\boldsymbol{S}\boldsymbol{q} + \frac{z}{1-z}\mathcal{T}(z)\boldsymbol{q} + \frac{1}{1-z}\boldsymbol{S}\boldsymbol{v}(0) + \mathcal{T}(z)\boldsymbol{v}(0)$$

- After inversion, $\boldsymbol{v}(n) = n\boldsymbol{S}\boldsymbol{q} + \mathcal{T}(1)\boldsymbol{q} + \boldsymbol{S}\boldsymbol{v}(0)$.
- If a column vector $\boldsymbol{g} = [g_i]$ is defined as $\boldsymbol{g} = \boldsymbol{S}\boldsymbol{q}$, then

$$\boldsymbol{v}(n) = n\boldsymbol{g} + \mathcal{T}(1)\boldsymbol{q} + \boldsymbol{S}\boldsymbol{v}(0). \tag{9}$$

Asymptotic Behavior: for long duration process

- Since any row of $\boldsymbol{S}$ is π, the steady state prob. vector of the MC, so all g_i are the same and $g_i = g = \sum_{i=1}^{N} \pi_i q_i$.
- Define $\boldsymbol{v} = \mathcal{T}(1)\boldsymbol{q} + \boldsymbol{S}\boldsymbol{v}(0)$, we have:

$$\boldsymbol{v}(n) \quad = \quad n\boldsymbol{g} + \boldsymbol{v} \quad \text{for large } n. \tag{10}$$

Example of asymptotic Behavior

For the toymaker's problem,

$$
\begin{aligned}
(\boldsymbol{I} - z\boldsymbol{P})^{-1} &= \frac{1}{1-z}\left[\begin{array}{cc} 4/9 & 5/9 \\ 4/9 & 5/9 \end{array}\right] + \frac{1}{1-\frac{1}{10}z}\left[\begin{array}{cc} 5/9 & -5/9 \\ -4/9 & 4/9 \end{array}\right] \\
&= \frac{1}{1-z}\boldsymbol{S} + \mathcal{T}(z)
\end{aligned}
$$

Since

$$
\boldsymbol{S} = \left[\begin{array}{cc} 4/9 & 5/9 \\ 4/9 & 5/9 \end{array}\right] \quad ; \quad \mathcal{T}(1) = \left[\begin{array}{cc} 50/81 & -50/81 \\ -40/81 & 40/81 \end{array}\right]
$$

$$
\boldsymbol{q} = \left[\begin{array}{c} 6 \\ -3 \end{array}\right] \quad ; \quad \boldsymbol{g} = \boldsymbol{S}\boldsymbol{q} = \left[\begin{array}{c} 1 \\ 1 \end{array}\right].
$$

By assumption, $\boldsymbol{v}(0) = 0$, then $\boldsymbol{v} = \mathcal{T}(1)\boldsymbol{q} = \left[\begin{array}{c} 50/9 \\ -40/9 \end{array}\right]$.

Therefore, we have $v_1(n) = n + \frac{50}{9}$ and $v_2(n) = n - \frac{40}{9}$.

Toymaker's Alternatives

- Suppose that the toymaker has other alternatives.
- If he has a successful toy, use advertising to decrease the chance that the toy will fall from favor.
- However, there is a *cost* to advertising and therefore the expected profit will generally be lower.
- If in state 1 and advertising is used, we have:

$$[p_{1,j}] = [0.8, 0.2] \quad [r_{1,j}] = [4, 4]$$

- In other words, for each state, the toymaker has to make a decision, advertise or not.

Toymaker's Alternatives

- In general we have policy 1 (no advertisement) and policy 2 (advertisement). Use superscript to represent policy.
- The transition probability matrices and rewards in state 1 (successful toy) are:

$$[p^1_{1,j}] = [0.5, 0.5], [r^1_{1,j}] = [9, 3];$$
$$[p^2_{1,j}] = [0.8, 0.2], [r^2_{1,j}] = [4, 4];$$

- The transition probability matrices and rewards in state 2 (unsuccessful toy) are:

$$[p^1_{2,j}] = [0.4, 0.6], [r^1_{2,j}] = [3, -7];$$
$$[p^2_{2,j}] = [0.7, 0.3], [r^2_{2,j}] = [1, -19];$$

Toymaker's Sequential Decision Process

- Suppose that the toymaker has n weeks remaining before his business will close down and n is the number of stages *remaining* in the process.
- The toymaker would like to know as a function of n and his present state, what alternative (policy) he should use to maximize the *total earning* over $n-$week period.
- Define $d_i(n)$ as the policy to use when the system is in state i and there are $n-$stages to go.
- Redefine $v_i^*(n)$ as the total expected return in n stages starting from state i if an **optimal** policy is used.

- We can formulate $v_i^*(n)$ as

$$v_i^*(n+1) = \max_k \sum_{j=1}^{N} p_{ij}^k \left[r_{ij}^k + v_j^*(n) \right] \qquad n = 0, 1, \ldots$$

- Based on the "Principle of Optimality", we have

$$v_i^*(n+1) = \max_k \left[q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j^*(n) \right] \qquad n = 0, 1, \ldots$$

In other words, we start from $n = 0$, then $n = 1$, and so on.

The numerical solution

- Assume $v_i^* = 0$ for $i = 1, 2$, we have:

$n =$	0	1	2	3	4	$\cdots$
$v_1(n)$	0	6	8.20	10.222	12.222	$\cdots$
$v_2(n)$	0	-3	-1.70	0.232	2.223	$\cdots$
$d_1(n)$	-	1	2	2	2	$\cdots$
$d_2(n)$	-	1	2	2	2	$\cdots$

Lessons learnt

- For $n \geq 2$ (greater than or equal to two weeks decision), it is better to do advertisement.
- For this problem, convergence seems to have taken place at $n = 2$. But for general problem, it is usually difficult to quantify.
- Some limitations of this value-iteration method:
  - What about *infinite stages*?
  - What about problems with many states (e.g., n is large) and many possible policies (e.g., k is large)?
  - What is the computational cost?

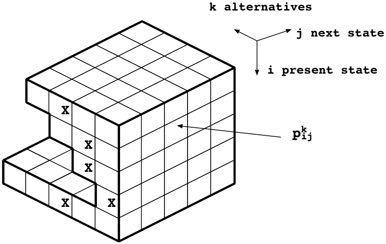
Preliminary

- From previous section, we know that the total expected earnings depend upon the total number of transitions (n), so the quantity can be unbounded.
- A more useful quantity is the **average earnings per unit time**.
- Assume we have an $N-$state Markov chain with one-step transition probability matrix $\boldsymbol{P} = [p_{ij}]$ and reward matrix $\boldsymbol{R} = [r_{ij}]$. Assume ergodic MC, we have the limiting state probabilities π_i for $i = 1, \ldots, N$, the gain g is

$$g = \sum_{i=1}^{N} \pi_i q_i; \quad \text{where} \quad q_i = \sum_{j=1}^{N} p_{ij} r_{ij} \;\; i = 1, \ldots, N.$$

A Possible five-state Markov Chain SDP

- Consider a MC with $N = 5$ states and $k = 5$ possible alternatives. It can be illustrated by



- X indicate the the chosen policy, we have $d = [3, 2, 2, 1, 3]$.
- Even for this small system, we have $4 \times 3 \times 2 \times 1 \times 5 = 120$ different policies.

Suppose we are operating under a given policy with a specific MC with rewards. Let $v_i(n)$ be the total expected reward that the system obtains in n transitions if it starts from state i. We have:

$$
\begin{aligned}
v_i(n) &= \sum_{j=1}^{N} p_{ij} r_{ij} + \sum_{j=1}^{N} p_{ij} v_j(n-1) \qquad n = 1, 2, \ldots \\
v_i(n) &= q_i + \sum_{j=1}^{N} p_{ij} v_j(n-1) \qquad n = 1, 2, \ldots \qquad (11)
\end{aligned}
$$

Previous, we derived the asymptotic expression of $\boldsymbol{v}(n)$ in Eq. (9) as

$$
v_i(n) = n \left(\sum_{i=1}^{N} \pi_i q_i \right) + v_i = ng + v_i \quad \text{for large } n. \qquad (12)
$$

For large number of transitions, we have:

$$
\begin{aligned}
ng + v_i &= q_i + \sum_{j=1}^{N} p_{ij}\left[(n-1)g + v_j\right] \qquad i = 1, ..., N \\
ng + v_i &= q_i + (n-1)g\sum_{j=1}^{N} p_{ij} + \sum_{j=1}^{N} p_{ij} v_j.
\end{aligned}
$$

Since $\sum_{j=1}^{N} p_{ij} = 1$, we have

$$
g + v_i = q_i + \sum_{j=1}^{N} p_{ij} v_j \qquad i = 1, \ldots, N. \tag{13}
$$

Now we have N linear simultaneous equations but $N+1$ unknown (v_i and g). To resolve this, set $v_N = 0$, and solve for other v_i and g. They will be called the relative values of the policy.

On Policy Improvement

- Given these relative values, we can use them to find a policy that has a higher gain than the original policy.
- If we had an optimal policy up to stage n, we could find the best alternative in the *ith* state at stage $n+1$ by

$$\arg\max_k q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j(n)$$

- For large n, we can perform substitution as

$$\arg\max_k q_i^k + \sum_{j=1}^{N} p_{ij}^k (ng + v_j) = \arg\max_k q_i^k + ng + \sum_{j=1}^{N} p_{ij}^k v_j.$$

- Since ng is independent of alternatives, we can maximize

$$\arg\max_k q_i^k \quad + \quad \sum_{j=1}^{N} p_{ij}^k v_j. \tag{14}$$

- We can use the relative values (v_j) from the value-determination operation for the policy that was used up to stage n and apply them to Eq. (14).
- In summary, the policy improvement is:
  - For each state i, find the alternative k which maximizes Eq. (14) using the relative values determined by the old policy.
  - The alternative k now becomes d_i the decision for state i.
  - A new policy has been determined when this procedure has been performed for every state.

The Policy Iteration Method

1. *Value-Determination Method:* use p_{ij} and q_i for a given policy to solve
$$g + v_i = q_i + \sum_{j=1}^{N} p_{ij} v_j \qquad i = 1, ..., N.$$
for all relative values of v_i and g by setting $v_N = 0$.
2. *Policy-Improvement Routine:* For each state i, find alternative k that maximizes
$$q_i^k + \sum_{j=1}^{N} p_{ij}^k v_j.$$
using v_i of the previous policy. The alternative k becomes the new decision for state i, q_i^k becomes q_i and p_{ij}^k becomes p_{ij}.
3. Test for convergence (check for d_i and g), if not, go back to step 1.

Toymaker's problem

For the toymaker we presented, we have policy 1 (no advertisement) and policy 2 (advertisement).

state i	alternative (k)	p_{i1}^k	p_{12}^k	r_{i1}^k	r_{i2}^k	q_i^k
1	no advertisement	0.5	0.5	9	3	6
1	advertisement	0.8	0.2	4	4	4
2	no advertisement	0.4	0.6	3	-7	-3
2	advertisement	0.7	0.3	1	-19	-5

Since there are two states and two alternatives, there are four policies, $(A, A), (\bar{A}, A), (A, \bar{A}), (\bar{A}, \bar{A})$, each with the associated transition probabilities and rewards. We want to find the policy that will maximize the average earning for indefinite rounds.

Start with policy-improvement

- Since we have no *a priori* knowledge about which policy is best, we set $v_1 = v_2 = 0$.
- Enter policy-improvement which will select an initial policy that maximizes the expected immediate reward for each state.
- Outcome is to select policy 1 for both states and we have

$$\boldsymbol{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{P} = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix} \quad \boldsymbol{q} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}$$

- Now we can enter the value-determination operation.

Value-determination operation

- Working equation: $g + v_i = q_i + \sum_{j=1}^{N} p_{ij} v_j$, for $i = 1, \ldots, N$.
- We have

$$g + v_1 = 6 + 0.5v_1 + 0.5v_2, \qquad g + v_2 = -3 + 0.4v_1 + 0.6v_2.$$

- Setting $v_2 = 0$ and solving the equation, we have

$$g = 1, \qquad v_1 = 10, \qquad v_2 = 0.$$

- Now enter policy-improvement routine.

Policy-improvement routine

State i	Alternative k	Test Quantity $q_i^k + \sum_{j=1}^N p_{ij}^k v_j$	
1	1	$6 + 0.5(10) + 0.5(0) = 11$	X
1	2	$4 + 0.8(10) + 0.2(0) = 12$	$\surd$
2	1	$-3 + 0.4(10) + 0.6(0) = 1$	X
2	2	$-5 + 0.7(10) + 0.3(0) = 2$	$\surd$

- Now we have a new policy, instead of $(\bar{A}, \bar{A})$, we have (A, A). Since the policy has not converged, enter value-determination.
- For this policy (A, A), we have

$$\boldsymbol{d} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad \boldsymbol{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} \quad \boldsymbol{q} = \begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

Value-determination operation

- We have

$$g + v_1 = 4 + 0.8v_1 + 0.2v_2, \qquad g + v_2 = -5 + 0.7v_1 + 0.3v_2.$$

- Setting $v_2 = 0$ and solving the equation, we have

$$g = 2, \qquad v_1 = 10, \qquad v_2 = 0.$$

- The gain of the policy (A, A) is thus twice that of the original policy, and the toymaker will earn **2 units per week** on the average, if he follows this policy.
- Enter the policy-improvement routine again to check for convergence, but since v_i didn't change, it converged and we stop.

The importance of discount factor β.

Working equation for SDP with discounting

- Let $v_i(n)$ be the **present value** of the total expected reward for a system in state i with n transitions before termination.

$$
\begin{aligned}
v_i(n) &= \sum_{j=1}^{N} p_{ij}\left[r_{ij} + \beta v_j(n-1)\right] \qquad i = 1,2,...,N, \;\; i = 1,2,... \\
&= q_i + \beta \sum_{j=1}^{N} p_{ij} v_j(n-1) \qquad i = 1,2,...,N. \; i = 1,2,... \quad (15)
\end{aligned}
$$

- The above equation also can represent the model of *uncertainty* (with probability β) of continuing another transition.

$Z-$transform of $\boldsymbol{v}(n)$

$$
\begin{aligned}
\boldsymbol{v}(n+1) &= \boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{v}(n) \\
z^{-1}\left[\boldsymbol{v}(z) - \boldsymbol{v}(0)\right] &= \frac{1}{1-z}\boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{v}(z) \\
\boldsymbol{v}(z) - \boldsymbol{v}(0) &= \frac{z}{1-z}\boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{v}(z) \\
\left(\boldsymbol{I} - \beta z \boldsymbol{P}\right)\boldsymbol{v}(z) &= \frac{z}{1-z}\boldsymbol{q} + \boldsymbol{v}(0) \\
\boldsymbol{v}(z) &= \frac{z}{1-z}\left(\boldsymbol{I} - \beta z \boldsymbol{P}\right)^{-1}\boldsymbol{q} + \left(\boldsymbol{I} - \beta z \boldsymbol{P}\right)^{-1}\boldsymbol{v}(0) \quad (16)
\end{aligned}
$$

Example

Using the toymaker's example, we have

$$\boldsymbol{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \quad \boldsymbol{P} = \begin{bmatrix} 1/2 & 1/2 \\ 2/5 & 3/5 \end{bmatrix}; \quad \boldsymbol{q} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}.$$

In short, he is **not** advertising and **not** doing research.

Also, there is a probability that he will go out of business after a week ($\beta = \frac{1}{2}$). If he goes out of business, his reward will be zero ($\boldsymbol{v}(0) = 0$).

What is the $\boldsymbol{v}(n)$?

Using Eq. (16), we have

$$\boldsymbol{v}(z) = \frac{z}{1-z}\left(\boldsymbol{I} - \beta z \boldsymbol{P}\right)^{-1}\boldsymbol{q} = \mathcal{H}(z)\boldsymbol{q}.$$

$$
\begin{aligned}
\left(\boldsymbol{I} - \frac{1}{2}z\boldsymbol{P}\right) &= \begin{bmatrix} 1-\frac{1}{4}z & -\frac{1}{4}z \\ -\frac{1}{5}z & 1-\frac{3}{10}z \end{bmatrix} \\
\left(\boldsymbol{I} - \frac{1}{2}z\boldsymbol{P}\right)^{-1} &= \begin{bmatrix} \frac{1-\frac{3}{10}z}{(1-\frac{1}{2}z)(1-\frac{1}{20}z)} & \frac{\frac{1}{4}z}{(1-\frac{1}{2}z)(1-\frac{1}{20}z)} \\ \frac{\frac{1}{5}z}{(1-\frac{1}{2}z)(1-\frac{1}{20}z)} & \frac{1-\frac{1}{4}z}{(1-\frac{1}{2}z)(1-\frac{1}{20}z)} \end{bmatrix} \\
\mathcal{H}(z) &= \begin{bmatrix} \frac{z(1-\frac{3}{10}z)}{(1-z)(1-\frac{1}{2}z)(1-\frac{1}{20}z)} & \frac{\frac{1}{4}z^2}{(1-z)(1-\frac{1}{2}z)(1-\frac{1}{20}z)} \\ \frac{\frac{1}{5}z^2}{(1-z)(1-\frac{1}{2}z)(1-\frac{1}{20}z)} & \frac{z(1-\frac{1}{4}z)}{(1-z)(1-\frac{1}{2}z)(1-\frac{1}{20}z)} \end{bmatrix}
\end{aligned}
$$

$$\mathcal{H}(z)=\frac{1}{1-z}\begin{bmatrix} \frac{28}{19} & \frac{10}{19} \\ \frac{8}{19} & \frac{30}{19} \end{bmatrix}+\frac{1}{1-\frac{1}{2}z}\begin{bmatrix} -\frac{8}{9} & -\frac{10}{9} \\ -\frac{8}{9} & -\frac{10}{9} \end{bmatrix}+\frac{1}{1-\frac{1}{20}z}\begin{bmatrix} -\frac{100}{171} & \frac{100}{171} \\ \frac{80}{171} & -\frac{80}{171} \end{bmatrix}$$

$$\boldsymbol{H}(n)=\begin{bmatrix} \frac{28}{19} & \frac{10}{19} \\ \frac{8}{19} & \frac{30}{19} \end{bmatrix}+\left(\frac{1}{2}\right)^n\begin{bmatrix} -\frac{8}{9} & -\frac{10}{9} \\ -\frac{8}{9} & -\frac{10}{9} \end{bmatrix}+\left(\frac{1}{20}\right)^n\begin{bmatrix} -\frac{100}{171} & \frac{100}{171} \\ \frac{80}{171} & -\frac{80}{171} \end{bmatrix}$$

Since $\boldsymbol{q}=\begin{bmatrix} 6 \\ -3 \end{bmatrix}$, we have

$$\boldsymbol{v}(n)=\begin{bmatrix} \frac{138}{19} \\ -\frac{42}{19} \end{bmatrix}+\left(\frac{1}{2}\right)^n\begin{bmatrix} -2 \\ -2 \end{bmatrix}+\left(\frac{1}{20}\right)^n\begin{bmatrix} -\frac{100}{19} \\ \frac{80}{9} \end{bmatrix}$$

Note that $n \to \infty$, $v_1(n) \to \frac{138}{19}$ and $v_2(n) \to -\frac{42}{19}$, which is **NOT** a function of n as the non-discount case.

What is the present value $\boldsymbol{v}(n)$ as $n \to \infty$?

From Eq. (15), we have $\boldsymbol{v}(n+1) = \boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{v}(n)$, hence

$$
\begin{aligned}
\boldsymbol{v}(1) &= \boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{v}(0) \\
\boldsymbol{v}(2) &= \boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{q} + \beta^2 \boldsymbol{P}^2 \boldsymbol{v}(0) \\
\boldsymbol{v}(3) &= \boldsymbol{q} + \beta \boldsymbol{P} \boldsymbol{q} + \beta^2 \boldsymbol{P}^2 \boldsymbol{q} + \beta^3 \boldsymbol{P}^3 \boldsymbol{v}(0) \\
\vdots &= \vdots \\
\boldsymbol{v}(n) &= \left[\sum_{j=0}^{n-1} (\beta \boldsymbol{P})^j \right] \boldsymbol{q} + \beta^n \boldsymbol{P}^n \boldsymbol{v}(0) \\
\boldsymbol{v} &= \lim_{n \to \infty} \boldsymbol{v}(n) = \left[\sum_{j=0}^{\infty} (\beta \boldsymbol{P})^j \right] \boldsymbol{q}
\end{aligned}
$$

What is the present value $\boldsymbol{v}(n)$ as $n \to \infty$?

Note that $\boldsymbol{v}(0) = \mathbf{0}$. Since $\boldsymbol{P}$ is a stochastic matrix, all its eigenvalues are less than or equal to 1, and the matrix $\beta\boldsymbol{P}$ has eigenvalues that are strictly less than 1 because $0 \leq \beta < 1$. We have

$$\boldsymbol{v} = \left[\sum_{j=0}^{\infty} (\beta\boldsymbol{P})^j\right] \boldsymbol{q} = (\boldsymbol{I} - \beta\boldsymbol{P})^{-1} \boldsymbol{q} \qquad (17)$$

Note: The above equation also provides a simple and efficient numerical method to compute $\boldsymbol{v}$.

Another way to solve $\boldsymbol{v}$

Direct Method

Another way to compute $\boldsymbol{v}_i$ is to solve N equations:

$$v_i = q_i + \beta \sum_{j=1}^{N} p_{ij} v_j \quad i = 1, 2, \ldots, N. \tag{18}$$

Consider the present value of the toymaker's problem with $\beta = \frac{1}{2}$ and

$$\boldsymbol{P} = \begin{bmatrix} 1/2 & 1/2 \\ 2/5 & 3/5 \end{bmatrix} \quad \boldsymbol{q} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}.$$

We have $v_1 = 6 + \frac{1}{4} v_1 + \frac{1}{4} v_2$ and $v_2 = -3 + \frac{1}{5} v_1 + \frac{3}{10} v_2$, with solution $v_1 = \frac{138}{19}$ and $v_2 = -\frac{42}{19}$.

Value Determination for infinite horizon

- Assume large n (or $n \rightarrow \infty$) and that $\boldsymbol{v}(0) = 0$.
- Evaluate the expected present reward for each state i using

$$v_i = q_i + \beta \sum_{j=1}^{N} p_{ij} v_j \quad i = 1, 2, \ldots, N. \tag{19}$$

for a given set of transition probabilities p_{ij} and the expected immediate reward q_i.

Policy-improvement

- The optimal policy is the one that has the highest present values in all states.
- If we had a policy that was optimal up to stage n, for state $n+1$, we should maximize $q_i^k + \beta \sum_{j=1}^{N} p_{ij} v_j(n)$ with respect to all alternative k in the i^{th} state.
- Since we are interested in the infinite horizon, we substitute v_j for $v_j(n)$, we have $q_i^k + \beta \sum_{j=1}^{N} p_{ij} v_j$.
- Suppose that the present value for an **arbitrary policy** have been determined, then a better policy is to maximize

$$q_i^k + \beta \sum_{j=1}^{N} p_{ij}^k v_j$$

  using v_i determined for the original policy. This k now becomes the new decision for the i^{th} state.

Iteration for SDP with Discounting

1. **Value-Determination Operation:** Use p_{ij} and q_i to solve th set of equations

$$v_i = q_i + \beta \sum_{j=1}^{N} p_{ij} v_j \qquad i = 1, 2, ..., N.$$

2. **Policy-Improvement Routing:** For each state i, find the alternative k^* that maximizes

$$q_i^k + \beta \sum_{j=1}^{N} p_{ij}^k v_j$$

   using the present values of v_j from the previous policy. Then k^* becomes the new decision for the *ith* state, $q_i^{k^*}$ becomes q_i and $p_{ij}^{k^*}$ becomes p_{ij}.

3. Check for convergence of policy. If not, go back to step 1, else halt.

Consider the toymaker's example with $\beta = 0.9$, we choose the initial policy that maximizes the expected immediate reward, we have

$$\boldsymbol{d} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \boldsymbol{P} = \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & 0.6 \end{bmatrix} \quad \boldsymbol{q} = \begin{bmatrix} 6 \\ -3 \end{bmatrix}$$

Using the *Value-Determination Operation*, we have

$$v_1 = 6 + 0.9(0.5v_1 + 0.5v_2) \qquad v_2 = -3 + 0.9(0.4v_1 + 0.6v_2)$$

The solution is $v_1 = 15.5$ and $v_2 = 5.6$.

Policy-improvement routine

State i	Alternative k	Value Test Quantity $q_i^k + \beta \sum_{j=1}^{N} p_{ij}^k v_j$	
1	1	$6 + 0.9[0.5(15.5) + 0.5(5.6)] = 15.5$	X
1	2	$4 + 0.9[0.8(15.5) + 0.2(5.6)] = 16.2$	$\surd$
2	1	$-3 + 0.9[0.4(15.5) + 0.6(5.6)] = 5.6$	X
2	2	$-5 + 0.9[0.7(15.5) + 0.3(5.6)] = 6.3$	$\surd$

- Now we have a new policy, instead of $(\bar{A}, \bar{A})$, we have (A, A). Since the policy has not converged, enter value-determination.
- For this policy (A, A), we have

$$\boldsymbol{d} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad \boldsymbol{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}, \quad \boldsymbol{q} = \begin{bmatrix} 4 \\ -5 \end{bmatrix}$$

Value-Determination Operation

Using the *Value-Determination Operation*, we have

$$v_1 = 4 + 0.9(0.8v_1 + 0.2v_2) \qquad v_2 = -5 + 0.9(0.7v_1 + 0.3v_2)$$

The solution is $v_1 = 22.2$ and $v_2 = 12.3$, which indicate a *signficant* increase in present values.

Policy-improvement routine

State i	Alternative k	Value Test Quantity $q_i^k + \beta \sum_{j=1}^{N} p_{ij}^k v_j$	
1	1	21.5	X
1	2	22.2	$\surd$
2	1	11.6	X
2	2	12.3	$\surd$

- The present value $v_1 = 22.2$ and $v_2 = 12.3$.
- Now we have the *same* policy (A, A). Since the policy remains the same, and the present values are the same. We can stop.